Capstone Project Report

A Clustering Analysis on Sydney Suburbs

1. Introduction

As the largest city in Australia with a 5 million population, Sydney renowned for its diverse culture and inclusive for numerous races. Sydney naturally becomes an ideal space for start-ups who want initial their own business and a large amount of opportunities are available. The diversity and complexity of Sydney's environment thus becomes a problem that makes the people hard to decide where to establish business.

On the other hand, clustering algorithms are well-developed in recent researches and studies to derive valuable patterns from unlabelled data in a novel view. This usually contribute to discover previous unknow segmentation of data and can be used for further research in certain area such as customer segmentation can be used to formulate sensible market strategy.

In addition, many suburbs of Sydney develop distinct and diverse cultures in the long history, forming their own typical community and supporting Sydney to become one of the most liveable cities around the world. Therefore, these suburbs are the basic elements of metropolitan Sydney and can be regard as individual entities that could be used as the input to analyse the business environment of Sydney.

Subsequently, using clustering algorithms to analysis and explore the similarity and dissimilarity between Sydney suburbs could lead to discover meaningful patterns, and these patterns can provides innovative insights for the corresponding government departments or companies who are interested in developing their business in Sydney.

2. Method

Detailed data of each suburb is obtained by saving the results of Foursquare venue query. Specifically, this query is achieved by conducting a explore search within the 500-radius circle area of the suburb centre. Then, category of venue will be extracted form the query result and frequency of different venues within the suburb will be calculated using one-hot method. This would create a suitable input for the K-Means Clustering. K-Means Clustering would be performed on the input data, and the outcome would be merged with the original data which has the geographical coordinates of each suburb. In this way, the result of Clustering would be visualized to present an conceive outcome for the audience.